GP-GAN: Gender Preserving GAN for Synthesizing Faces from Landmarks

Xing Di, Vishwanath A. Sindagi, Vishal M. Patel

Rutgers, The State University of New Jersey

94 Brett Road, Piscataway, NJ 08854

{xing.di, vishwanath.sindagi, vishal.m.patel}@rutgers.edu

Abstract—**Facial landmarks constitute the most compressed representation of faces and are known to preserve information such as pose, gender and facial structure present in the faces. Several works exist that attempt to perform high-level face-related analysis tasks based on landmarks alone without the aid of face images. In contrast, in this work, an attempt is made to tackle the inverse problem of synthesizing faces from their respective landmarks. The primary aim of this work is to demonstrate that information preserved by landmarks (gender in particular) can be further accentuated by leveraging generative models to synthesize corresponding faces. Though the problem is particularly challenging due to its ill-posed nature, we believe that successful synthesis will enable several applications such as boosting performance of high-level face related tasks using landmark points and performing dataset augmentation. To this end, a novel face-synthesis method known as Gender Preserving Generative Adversarial Network (GP-GAN) that is guided by adversarial loss, perceptual loss and a gender preserving loss is presented. Further, we propose a novel generator sub-network UDeNet for GP-GAN that leverages advantages of U-Net and DenseNet architectures. Extensive experiments and comparison with recent methods are performed to verify the effectiveness of the proposed method. Our code is available at: https://github.com/DetionDX/GP-GAN-Gender-Preserving-GAN-for-Synthesizing-Faces-from-Landmarks**

I. INTRODUCTION

Facial landmarks can be regarded as the most compressed representation of a face due to the fact that very few number of points are required to capture the landmark locations. In spite of the incredibly low number of keypoints, they are known to preserve important information about the face such as pose, gender [4] and structure [34], [28], [33]. Success of facial analysis tasks using just landmark keypoints is essential from the perspective of memory management and information privacy. Considering that size of landmarks is an order of magnitude smaller as compared to the image size, it will result in significant savings in terms of memory. Essentially, we are now able to store only landmark key points and throw away face image for a particular application. In addition, landmark information can be safely stored, transported, and distributed without potential violation of human privacy and confidentiality. Motivated by these reasons, it would be interesting to understand how landmarks can be exploited for performing high-level facial analysis tasks in the absence of corresponding face images.

Several researchers have demonstrated that facial landmarks can be used in many face analysis tasks such as face
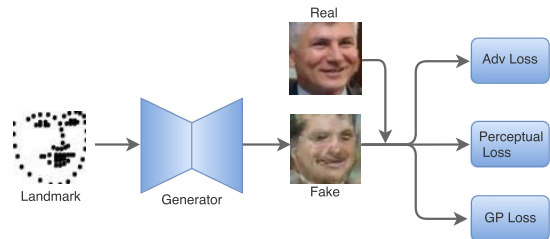

Fig. 1. Overview of the proposed GP-GAN method for synthesizing faces from landmarks. In addition to adversarial loss function, the generator sub-network is guided by a perceptual loss and a gender preserving loss.

recognition [5], [19], [20], facial attribute inference [41], age estimation [34], gender recognition [4] and expression analysis [30]. However, these methods operate on a small set of keypoints due to which their performance is severely limited. To overcome this problem, we propose a novel solution that involves synthesis of faces from landmark points using the recently popular generative models [7], [43], [2], [42], [38], [32], [37]. While, several methods [41], [31], [14], [29] have been proposed in the literature for landmark detection, the inverse problem of synthesizing faces from their corresponding landmarks is a largely unexplored problem. We believe that using synthesized faces will result in better recognition performances as they leverage the capabilities of generative models to accentuate information present in landmarks. Apart from their use in high-level facial analysis tasks, these generative methods can be used to create virtually unlimited stochastic samples by conditioning on both landmarks and a stochastic noise vector enabling us to augment existing datasets for large scale learning [3].

In this work, generative models are exploited to synthesize faces from landmarks in an attempt to accentuate information (gender in particular) present in the landmarks. Cao *et al.*[4] specifically address the question if facial metrology can be used to predict gender and they further go on to demonstrate that gender recognition using landmarks achieves reasonable performance. This is remarkable considering the fact that only 68 keypoints are used to predict gender of the face represented by these keypoints. However, generating faces from landmarks will enable us to achieve further improvement in performance as this process will leverage generative models to learn the distribution of landmarks and their mappings to the respective faces. While recognition of other attributes like ethnicity, pose,

identity, etc. can all be improved, in this work, we specifically focus on the gender attribute. To this end, we propose Gender Preserving Generative Adversarial Network (GP-GAN) to generate faces from their respective landmarks (as shown in Fig. 1). To further enhance the network's performance, it is guided by perceptual loss and a gender preserving loss in addition to adversarial loss. To summarize, following are the key contributions of this work:

- To the best of our knowledge, this is the first attempt to generate faces from landmark keypoints while preserving gender information.
- A GAN-based framework guided by perceptual and gender preserving loss is proposed. The generator is constructed using a novel combination of UNet [23] and DenseNet [8], which we call it as UDeNet.
- Detailed experiments are conducted to demonstrate the improvements in gender recognition obtained from synthesized images using the proposed method.

II. RELATED WORK

In contrast to landmark detection methods [24], [29], [14], [1] , we focus on the inverse problem of synthesizing or generating faces from landmark keypoints which is a relatively unexplored problem. To this end, recently popular generative models are explored in this work. Among these methods, we specifically study Generative Adversarial Network (GAN)[7], [43], [2], [42] and Variational Auto-encoder (VAE) [22], [13].

VAEs are powerful generative models that use deep networks to describe distribution of observed and latent variables. A VAE consists of two networks, with one network encoding a data sample to a latent representation and the other network decoding latent representation back to data space. VAE regularizes the encoder by imposing a prior over the latent distribution. Conditional VAE (CVAE) [27] [35] [6] is an extension of VAE that models latent variables and data, both conditioned on side information such as a part or label of the image. GANs [7] are another class of generative models that are used to synthesize realistic images by effectively learning the distribution of training images. Recently, several variants based on this game theoretic approach have been proposed for image-to-image translation tasks. Isola *et al.*[11] proposed Conditional GANs [17] for several tasks such as labels to street scenes, labels to facades, image colorization, etc. In an another variant, Zhu *et al.*[43] proposed CycleGAN that learns image-to-image translation in an unsupervised fashion. Berthelot *et al.*[2] proposed a new method for training auto-encoder based GANs that is relatively more stable. Their method is paired with a loss inspired by Wasserstein distance. Some of the other applications of GANs include image de-hazing [39], crowd counting [26], and image de-raining [38].

III. PROPOSED METHOD

Given an application where only facial landmarks are available, we explore how to leverage information preserved in these keypoints. To this end, we propose to model the joint distribution of facial landmarks and corresponding face images [1] using generative modeling. Inspired by the success of GANs [7], we explore adversarial networks in this work for synthesizing faces from landmark keypoints. GANs, motivated by game theory, consist of two competing networks: generator G and discriminator D. The goal of GAN is to train G to produce samples from training distribution such that the synthesized samples are indistinguishable from actual distribution by discriminator D. Conditional GAN is another variant where the generator is conditioned on additional variables such as discrete labels [17], text [21] and images [11]. The objective function of a conditional GAN is defined as follows

$$L_{cGAN}(G, D) = E_{x,y \sim Pdata(x,y)}[\log D(x,y)] + \\ E_{x \sim Pdata(x), z \sim p_z(z)}[\log(1 - D(x, G(x, z)))], \quad (1)$$

where y, the output image, and x, the observed image, are sampled from distribution $Pdata(x, y)$ and they are distinguished by the discriminator, D. While for the generated fake $G(x, z)$ sampled from distributions $x \sim Pdata(x), z \sim p_z(z)$ would like to fool D.

As shown in Fig. 2, the proposed network consists of a generator sub-network G (based on U-net [23] and DenseNet [8] architecture) conditioned on a facial landmark image and a patch-based discriminator sub-network D. G takes landmark as input and attempts to generate corresponding face image, while D attempts to distinguish between real and synthesized images. The two sub-networks are trained iteratively. In addition to the adversarial loss, we propose to guide the generator using three other loss functions: perceptual loss based on VGG-16 architecture [25], gender preserving loss and L_1 reconstruction error.

A. Generator

Deeper networks are known to better capture high-level concepts, however, the vanishing gradient problem affects convergence rate as well as the quality of convergence. Several works have been developed to overcome this issue among which U-Net [23] and DenseNet [8] are of particular interest. While U-Net incorporates longer skip connections to preserve low-level features, DenseNet employs short range connections within micro-blocks resulting in maximum information flow between layers in addition to an efficient network. Motivated by these two methods, we propose UDeNet for the generator sub-network G in which, the U-Net architecture is seamlessly integrated into the DenseNet network in order to leverage advantages of both the methods. This novel combination enables more efficient learning and improved convergence quality.

A set of 3 dense-blocks (along with transition blocks) are stacked in the front, followed by a set of 5 dense-block layers (transition blocks). The initial set of dense-blocks are composed of 6 bottleneck layers. For efficient training and better convergence, symmetric skip connections are involved into the generator sub-network, similar to [16]. Details regarding the number

---

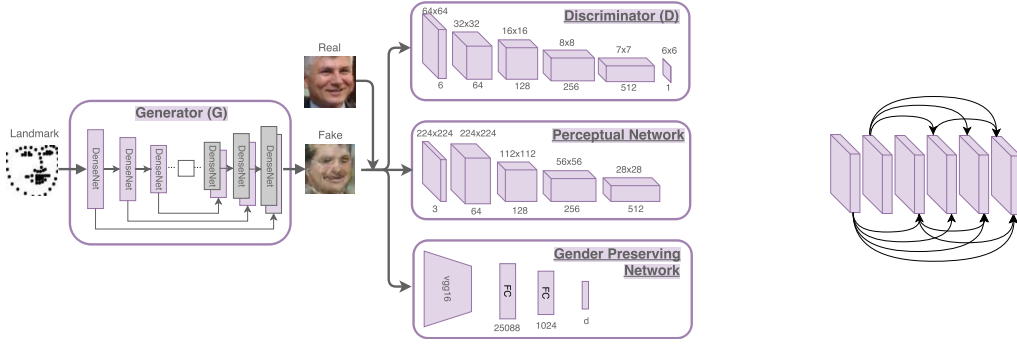[1] Face images are available only during training

Fig. 2. Architecture of the proposed GP-GAN framework. Left: Generator (G) synthesizes face image from landmarks and is based on UNet and DenseNet architecture. D is a patch-based discriminator that is trained to distinguish between real/fake face images and it is responsible for providing adversarial feedback to G. G is also guided by a perceptual loss (based on VGG-16 architecture) and a gender-preserving loss. Right: Dense-block used in generator G.

of channels for each convolutional layer are as follows:
C(64)-M(64)-D(256)-T(128)-D(512)-T(256)-
D(1024)-T(512)-D(1024)-DT(256)-D(512)-DT(128)-
D(256)-DT(64)-D(64)-D(32)-D(32)-DT(16)-C(3),
where C(K) is a set of K-channel convolutional layers followed by batch normalization and ReLU activation. M is max-pooling layer. D(K) is the dense-block layer with K-channel output, T(K) is transition layer with K-channel output for downsampling. DT(K) is similar to T(K) except for transposed convolutional layer instead of convolutional layer for upsampling.

B. Discriminator

Motivated by [11], patch-based discriminator D is used and it is trained iteratively along with G. The primary goal of D is to learn to discriminate between real and synthesized samples. This information is backpropagated into G so that it generates samples that are as realistic as possible. Additionally, patch-based discriminator ensures preserving of high-frequency details which are usually lost when only L1 loss is used. All the convolutional layers in D have a filter size of 4×4. Details regarding the number of channels for each convolutional layer are specified in Fig. 2.

C. Objective function

The network parameters are learned by minimizing the following objective function:

$$L = L_A + \lambda_P L_P + \lambda_C L_C + \lambda_1 L_1, \qquad (2)$$

where L_A is the adversarial loss, L_P is the perceptual loss, L_C is the gender preserving loss and L_1 is the loss based on L_1-norm between the target and reconstructed image, λ_P, λ_C and λ_1 are weights respectively for perceptual loss, gender preserving loss and L_1 loss.

Adversarial loss: Adversarial loss is based primarily on the discriminator sub-network D. Given a set of N synthesized faces, $\{\hat{x}_i\}_{i=1}^N$, the entropy loss from D that is used to learn the parameters of G is defined as:

$$L_A = -\frac{1}{N} \sum_{i=1}^N log(D(\hat{x}_i)), \qquad (3)$$

Perceptual loss: Johnson *et al.*[12] introduced the perceptual loss function for style transfer and super-resolution. Instead of relying only on L_1 or L_2 reconstruction error, they learn the network parameters using errors between high-level image feature representations extracted from a pre-trained convolutional neural network. Similar to their work, pre-trained VGG-16 [25] network is used to extract high-level features (conv4_3 layers) and the L_1 distance between these features of real and fake images is used to guide the generator G. The perceptual loss function is defined as:

$$L_P = ||V(\hat{x}) - V(x)||_1, \qquad (4)$$

where, x and $\hat{x}$ indicate real and fake images, respectively and V is a particular layer of the VGG-16 network.

Gender preserving loss: Inspired largely by the perceptual loss, we define a gender preserving loss. As indicated by the name, this function measures the error in terms of gender attribute of the synthesized image as compared to that of real image. It is defined as:

$$L_C = -\frac{1}{N} \sum_i (C(x_i) \log(C(\hat{x}_i)) + (1 - C(x_i)) \log(C(\hat{x}_i)),$$

where C represents a pre-trained gender classification network. In this work, C is constructed using the standard VGG-16 network in which, the convolutional layers are retained and the fully connected layers are replaced by a new set of layers as shown in Fig. 2. This network is trained by minimizing the standard binary cross entropy error.

L1 loss: L1 loss measures the reconstruction error between the synthesized face image and the corresponding real image and is defined as

$$L_1 = ||G(\hat{x}) - x||_1 \qquad (5)$$

TABLE I
QUANTITATIVE COMPARISON OF GENDER RECOGNITION ACCURACY (%) FOR VARIOUS METHODS.

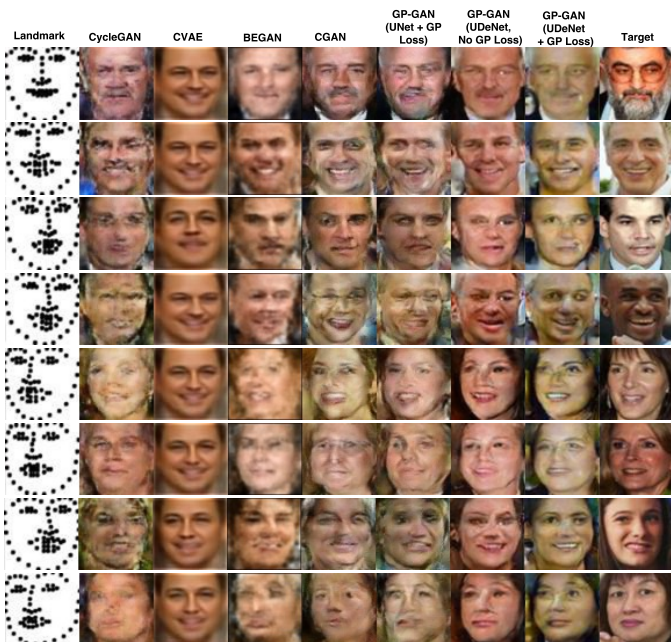|	LM (D)	LM (A)	CycleGAN	CVAE	BEGAN	CGAN	GP-GAN (UNet+GP-Loss)	GP-GAN (UDeNet,No GP Loss)	GP-GAN (UDeNet+GP-Loss)
LFW	78.0 ± 1.9	79.8 ± 2.4	81.8 ± 1.1	80.3 ± 2.0	84.4 ± 1.9	86.3 ± 2.5	91.1 ± 1.1	91.7 ± 1.6	93.1 ± 1.2
CASIA	61.0 ± 11.8	61.7 ± 13.6	64.8 ± 3.3	62.0 ± 4.1	67.8 ± 5.0	70.4 ± 5.5	73.2 ± 3.9	76.7 ± 4.3	78.4 ± 4.1

Fig. 3. Sample qualitative results of synthesis experiments from LFW dataset. The proposed method GP-GAN (UDeNet + GP Loss) achieves more realistic synthesis compared to the other methods (CycleGAN, CVAE, BEGAN, CGAN) and the baseline methods from the ablation study: GP-GAN (UNet+GP Loss), GP-GAN (UDeNet+ No GP Loss).
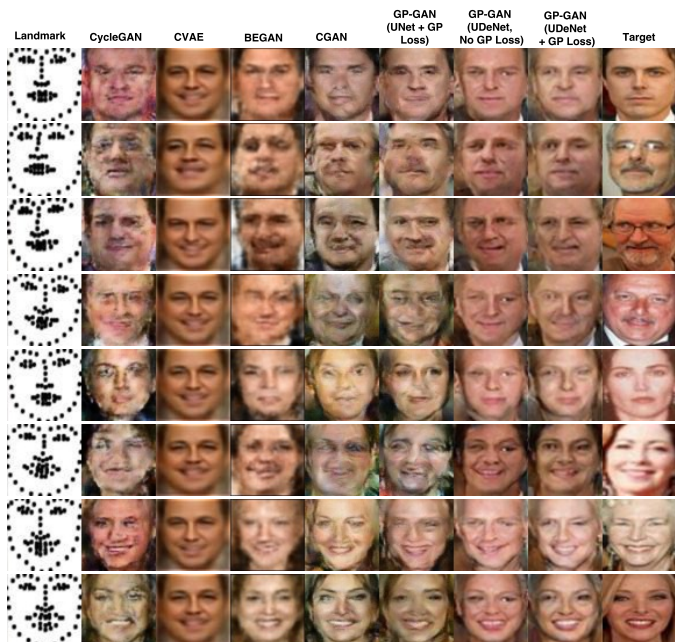

Fig. 4. Sample qualitative results of synthesis experiments from CASIA WebFace dataset. The proposed method GP-GAN (UDeNet + GP Loss) achieves more realistic synthesis compared to the other methods (CycleGAN, CVAE, BEGAN, CGAN) and the baseline methods from the ablation study: GP-GAN (UNet+GP Loss), GP-GAN (UDeNet+ No GP Loss).

IV. EXPERIMENTS AND EVALUATIONS

In this section, experimental settings and evaluation of the proposed method are discussed in detail. We present the qualitative and quantitative results of the synthesis experiment. The quantitative performance is measured using gender recognition rates. Results are compared with four state-of-the-art generative models: Conditional GAN [11], Cycle GAN [43], CVAE [27] [35] and adopted BEGAN[2] in addition to two baseline methods (a) GP-GAN using U-Net generator with GP-Loss, and (b) GP-GAN using UDeNet generator without GP-Loss. The baseline comparisons are performed to demonstrate the improvements achieved by the gender preserving loss and UDeNet components. Also, we demonstrate that the use of synthesis using GP-GAN accentuates gender information present in landmarks by comparing gender recognition rates with methods that directly compute these rates from landmark points [4]. Furthermore, we conduct an experiment to evaluate the data augmentation capabilities of the synthesis method.

A. Preprocessing and training details

Prior to preforming these experiments, all images in both datasets are fed through a pre-processing pipeline. First,

[2]https://github.com/taey16/pix2pixBEGAN.pytorch

MTCNN [40] is employed for detecting face bounding boxes which are further used to crop the faces followed by landmark key point detection using TCDCN algorithm [41]. Pairs of these detected landmarks and faces are used for training the proposed method. Since we consider this problem as an image-to-image translation, the input landmark is encoded using a heatmap (similar to [15]) as shown in Fig. 1 which is a created by imposing a 2D Gaussian with standard deviation of 0.2 at every landmark location on a blank image like could counting work [26]. Note that the cropped face images are resized to 64×64.

The proposed network is trained on a single TitanX GPU for approximately 10 hours (200 epochs). A learning rate of 2×10^{-4} is used for G and D. For perceptual network, the input images are resized to a size of 224×224. The learning rate is decayed by a factor of 2×10^{-6} for every epoch after 100 epochs. The weights λ_A, λ_P and λ_C are set equal to 100, 1 and 1, respectively.

For learning the parameters of the proposed method and baselines, training set from the LFW official deep funneling aligned dataset [10][9] is used. It contains 5749 identities, and 13233 images. The official training, validating and testing View 1 was used for this experiment. After detection and crop procedure, we are left with 3757 images in the training

set and 1615 images in the test set. The trained network is evaluated on the LFW test set and a subset of CASIA-Webface dataset [36]. The test subset for CASIA-Webface is constructed by randomly selecting 1000 male and 1000 female face images. Note that, in order to demonstrate the generalization performance, the proposed network is trained using only the LFW training set and evaluated on the LFW test set and the CASIA-Webface dataset.

B. Results

Fig. 3 and Fig. 4 show sample results of reconstruction using various methods on the LFW and CASIA datasets, respectively. The landmark image is used as the input for all the methods except CVAE [27] [35]. For CVAE, the inputs are original image and normalized landmark locations as the attributes. It can be clearly observed that Conditional GAN [11], Cycle GAN [43] and BEGAN [2] are unable to reconstruct visually coherent faces. Though CVAE is able to generate visually appropriate faces, they fail to preserve the gender information. Since their network implements an auto-encoder like architecture and uses pixel-wise Euclidean measure, the output is often blurry, due to which gender classification becomes very difficult. GP-GAN using UDeNet generator without GP-Loss is able to generate perceptually better results as compared to GP-GAN using UNet generator with GP-Loss demonstrating the superior performance obtained using the novel combination of UNet and DenseNet architectures. The proposed method GP-GAN (UDeNet and GP-Loss) outperforms all existing and baseline methods. It may be argued that identity information is lost during the reconstruction process, however, note that the goal of the proposed method is not to capture the exact mapping between landmarks and corresponding faces. Instead, the idea is to explore generation of visually coherent faces from landmark keypoints which can further assist in data augmentation and other tasks.

As discussed earlier, the quantitative performance is measured in terms of gender recognition rates and it is shown in Table I. Gender recognition rates for the synthesized are calculated using the LBP features [18] and a linear SVM classifier that is trained using the LFW training set, whereas the recognition rates for landmarks, LM(D) and LM(A), are calculated using the distance and angle methods described in [4]. Note that the gender recognition is performed based only on landmark keypoints considering that the corresponding face images are unavailable and hence recent state-of-the-art gender recognition methods cannot be used for comparison as they operate on actual face images rather than only on facial landmarks. Similar to the observations made using visual comparisons, it can be found from the quantitative results that, gender recognition rates improve in general using the generative models as compared to the landmark-based methods.

With respect to the baseline comparisons, it can be observed that GP-GAN using UDeNet generator without GP-Loss outperforms GP-GAN using UNet generator with GP-Loss in
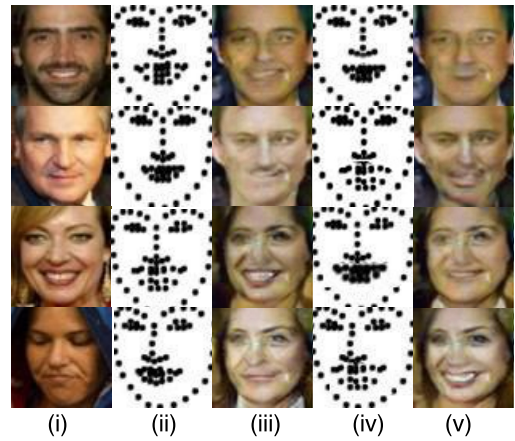

Fig. 5. Results of experiment for dataset augmentation where landmark corresponding to a face is modified and used for synthesis. We are able to generate new samples while preserving gender information. (i) Original face image. (ii) Landmark corresponding to original face. (iii) Synthesized face from original landmark. (iv) Landmark obtained after manipulating original landmark. (v) Synthesized face image using manipulated landmark.

spite of the fact that GP-Loss is not used, thus indicating the effectiveness of UDeNet architecture. Furthermore, the proposed method GP-GAN (UDeNet with GP-Loss) outperforms all existing baseline methods by a large margin in terms of gender recognition rates. This indicates that the proposed synthesis method can be used to generate face images from just facial landmarks while retaining gender information present in these landmarks.

In addition, we conducted a face synthesis experiment to verify if the proposed method can be used for data augmentation. In this experiment, we manipulate the landmark of a face (for instance, modify mouth open to mouth close) and use this landmark to synthesize a face using generator G. Sample results for this experiment are shown in Fig 5. It can be seen that, the generator G is able to synthesize realistic faces from the modified landmarks while reflecting this modification in the synthesized face. Additionally, the gender attribute is also retained. Based on these experiments, we can conclude that the proposed method is able successfully generate face samples which can be used for data augmentation for other facial analysis tasks.

V. CONCLUSION

We explored the problem of synthesizing faces from landmarks points using the recently introduced generative models. The aim of this project was to demonstrate that information (especially gender) present in the landmark keypoints can be accentuated using synthesis models while generating realistic images. The proposed network is based on the generative adversarial networks and is guided by perceptual loss and a novel gender preserving loss. Further, we propose a novel generator based on UNet and DenseNet architectures. Evaluations are performed on two popular datasets, LFW and CASIA-Webface, and the results are compared with recent

state-of-the-art generative methods. It is clearly demonstrated that the proposed method achieves significant improvements in terms of visual quality and gender recognition. Additionally, we conducted a face synthesis experiment to demonstrate that the proposed generative method can be used as a data augmentation technique.

ACKNOWLEDGMENTS

REFERENCES

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *arXiv preprint arXiv:1612.05424*, 2016.

[4] D. Cao, C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross. Can facial metrology predict gender? In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE, 2011.

[5] J. C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[6] X. Di and V. M. Patel. Face synthesis from visual attributes via sketch using conditional vaes and gans. *arXiv preprint arXiv:1801.00077*, 2017.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller. Learning to align from scratch. In *NIPS*, 2012.

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[13] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] A. Kumar, R. Ranjan, V. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*, 2016.

[15] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NIPS*, 2017.

[16] X.-J. Mao, C. Shen, and Y.-B. Yang. Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections. *arXiv preprint*, 2016.

[17] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[19] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2017.

[20] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, Jan 2018.

[21] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[22] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backprop-agation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[24] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] V. A. Sindagi and V. M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *IEEE International Conference on Computer Vision*, 2017.

[27] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[28] Q. Sun, L. Ma, S. J. Oh, L. V. Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. In *CVPR*, 2018.

[29] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[30] S. Taheri, P. Turaga, and R. Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 306–313. IEEE, 2011.

[31] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE, 2010.

[32] L. Wang, V. A. Sindagi, and V. M. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. *CoRR*, abs/1710.10182, 2017.

[33] W. Wang, X. Alameda-Pineda, D. Xu, E. Ricci, and N. Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *CVPR*, 2018.

[34] T. Wu, P. Turaga, and R. Chellappa. Age estimation and face verification across aging using landmarks. *IEEE Transactions on Information Forensics and Security*, 7(6):1780–1788, 2012.

[35] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[36] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[37] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. *arXiv preprint arXiv:1803.08396*, 2018.

[38] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017.

[39] H. Zhang, V. Sindagi, and V. M. Patel. Joint transmission map estimation and dehazing using deep networks. *arXiv preprint arXiv:1708.00581*, 2017.

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[41] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108. Springer, 2014.

[42] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*

arXiv:1703.10593, 2017.